Enhanced Event-based Dense Stereo via Cross-Sensor Knowledge Distillation

Supplementary Material

1. Experimental Details

Datasets. In this work, we use the MVSEC [5] and DSEC [3] datasets. For MVSEC, we use indoor flying, which was captured by a drone in a room with normal lighting. For DSEC, we approximately map the intensity images to the corresponding positions of the events with the provided camera matrix. After that, we overlay the events on the corresponding intensity images for visualization and hand it over to professionals for inspection to ensure alignment. We re-split the train set and the test set of DSEC, as shown in Table 1.

Set	Sequences	Names
Train	Zurich City	zurich_city_04_a, zurich_city_04_b, zurich_city_04_c, zurich_city_05_a, zurich_city_05_b, zurich_city_06_a, zurich_city_07_a, zurich_city_08_a.
	Interlaken	interlaken_00_c, interlaken_00_d, interlaken_00_e.
Test	Zurich City	zurich_city_11_a, zurich_city_11_b, zurich_city_11_c.
	Interlaken	interlaken_00_f, interlaken_00_g.

Table 1. Detail of the DSEC dataset splits.

Implementation Details. The workflow of **IED^{2}S** can be summarized as Algorithm 1. The intensity branch is pre-trained with only disparity loss for 100 epochs, while the event branch is trained via cross-sensor distillation with the full loss for 64 epochs. For the parameters β_1 and β_2 in Eq. (5), we set them to 1 and 0.6 respectively.

```
Algorithm 1: Process in intensity-to-event distillation for dense stereo matching (IED$^2$S).
Data: The events of the left and right event cameras $\{E_i^L\}_{i=1}^M$ and $\{E_i^R\}_{i=1}^M$, corresponding intensity images $\{I_i^L\}_{i=1}^M$ and $\{I_i^R\}_{i=1}^M$, and the ground truth disparity $\{D_i\}_{i=1}^M$, where M is a number of data samples.
Stage 1: Training Intensity Branch: Taking stereo intensity images set $\{I_i^L\}_{i=1}^M$ and $\{I_i^R\}_{i=1}^M$ as input, image encoder produce the intensity cost volume $\{F_i^{inten}\}_{i=1}^M$, which are then fed into disparity regression and obtain prediction disparities supervised by ground-truth labels $\{D_i\}_{i=1}^M$.
Stage 2: Training Event Branch via Cross-Sensor Distillation: Taking stereo intensity images and events as input, intensity and event encoder produce the intensity cost volume $\{F_i^{inten}\}_{i=1}^M$ and event cost volume $\{F_i^{event}\}_{i=1}^M$, which are then fed into disparity regression and obtain prediction disparities $\{d_i^{inten}\}_{i=1}^M$ and $\{d_i^{event}\}_{i=1}^M$ supervised by ground-truth labels $\{D_i\}_{i=1}^M$. Cross-sensor distillation is conducted through novel design between cost volume $\{F_i^{inten}\}_{i=1}^M$ and $\{F_i^{event}\}_{i=1}^M$.
```

2. Feature Extractor

For the feature extractor, we follow the design of a well-performing stereo matching network [1] and use its feature extractor to extract features for events as well as intensity images. Since the input channels of events and intensity images are different, the corresponding feature extractors have the same structure except for the first convolutional layer. The feature extractors for left events and right events share weights, and similarly, the feature extractors for left intensity images and right intensity images share weights. After training, only the feature extractor for the event branch is retained.

3. More Qualitative Results

3.1. The Effectiveness of Cross-Sensor Distillation (CSD)

We provide qualitative results for our proposed **CSD** in Fig. 1. We have done quantitative ablation studies, and the qualitative results will clearly demonstrate the effectiveness. Compared with the model without CSD, the proposed method effectively distills the rich semantic information in the intensity image to the event branch. The predicted result effectively separates the object from the background and achieves clear disparity estimation.

3.2. Additional Evaluation on the MVSEC Dataset

We provide more qualitative comparisons on the MVSEC datasets in the Fig. 2. Our method surpasses event-based stereo method. In other words, using both modalities together during training is effective for stereo performance. Furthermore, the results show that although event data can reflect edge and boundary information, due to the sparsity of the event data itself, dense disparity cannot be accurately obtained through this modality alone. When images with dense information are used together, stereo matching shows effective results. In addition, our model can clearly distinguish objects from the background by spatial location correlation of events and intensity images. One problem with intensity-based stereo matching is the edges at depth discontinuities. Our method fully solves the problem of flat edges because we do not input intensity images during the inference phase.

4. Discussions

4.1. Value of this Work

In this paper, we propose an intensity-to-event distillation for dense stereo matching. We demonstrate the effectiveness of the proposed components from qualitative and quantitative experimental results. Our method achieves significant performance improvements over previous state-of-the-art event-based methods [2]. This work focuses on stereo matching, but our framework is generally applicable to tasks that use both intensity images and events during the training phase. The well-designed multi-level distillation strategy can serve as a **template** for other researchers to tackle cross-modality challenges. In future work, we will extend our approach to other tasks that require complementary information from two modalities, such as optical flow estimation, object detection, etc.

The cross-modal distillation proposed in this work enables the event branch to learn general knowledge and has good generalization, which makes it possible to achieve dense stereo matching by only collecting events in practical applications.

4.2. The Motivation for Using Events

The hardware advantages of event cameras, with very low latency and high dynamic range, make them essentially unaffected by motion blur and suitable for extreme lighting scenarios. They are very suitable for certain specific driving scenarios and can enrich the information missing due to the defects of frame-based cameras as a complementary source. On the other hand, event cameras mainly focus on the edge information of objects. This makes them an ideal tool for stereo matching. However, they are still sparse in the overall area, so we carefully designed cross-sensor distillation to make up for this shortcoming and make full use of the complementary information of the two modalities.

4.3. Motivation for Distillation Design

The cost volume encodes the probability of each pixel's disparity. Distilling it allows the event branch to learn rich spatial information. We avoid distilling features directly to reduce complexity and prevent imbalance from applying losses on both left and right features. Taking the above reasons into consideration, we choose to distill the information contained in the cost volume. Specifically, we distill along the disparity dimension to capture information across all disparity ranges, producing a 3D tensor that encodes the information 061 belonging to that disparity value.

References

[1] Jia-Ren Chang and Yong-Sheng Chen. Pyramid stereo matching network. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5410–5418, 2018. 1

[2] Hoonhee Cho, Jae-Young Kang, and Kuk-Jin Yoon. Temporal event stereo via joint learning with stereoscopic flow. In *Proceedings of the European Conference on Computer Vision*, pages 294–314. Springer, 2025. 2, 4

[3] Mathias Gehrig, Willem Aarents, Daniel Gehrig, and Davide Scaramuzza. Dsec: A stereo event camera dataset for driving scenarios. *IEEE Robotics and Automation Letters*, 6(3):4947–4954, 2021. 1

[4] Stepan Tulyakov, Francois Fleuret, Martin Kiefel, Peter Gehler, and Michael Hirsch. Learning an event sequence embedding for dense event-based deep stereo. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1527–1537, 2019. 4

[5] Alex Zihao Zhu, Dinesh Thakur, Tolga Özaslan, Bernd Pfrommer, Vijay Kumar, and Kostas Daniilidis. The multivehicle stereo event camera dataset: An event camera dataset for 3d perception. *IEEE Robotics and Automation Letters*, 3(3): 2032–2039, 2018. 1

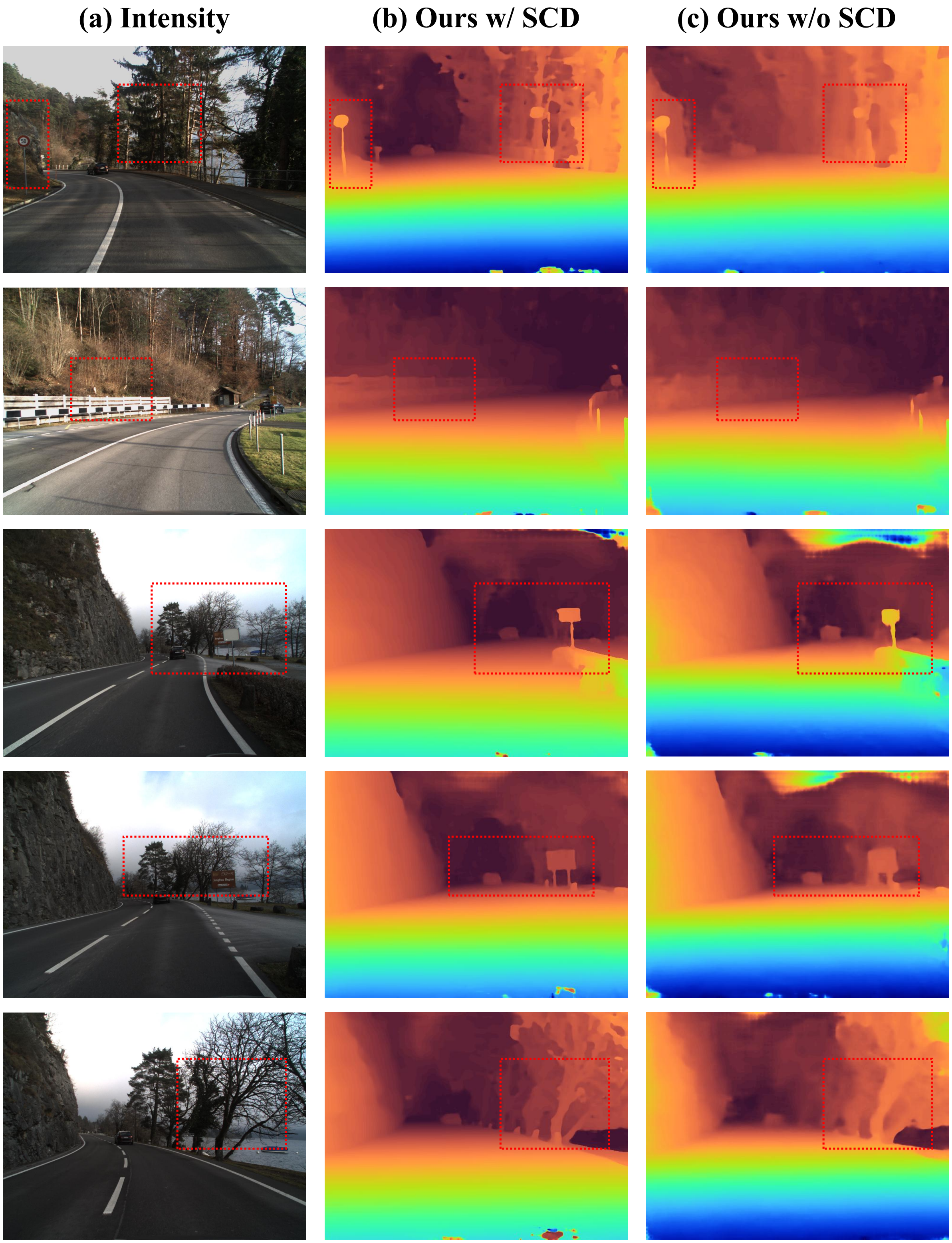

Figure 1. Qualitative results for our proposed **CSD** in the DSEC dataset. Detailed disparity information is framed by the red box for comparison.

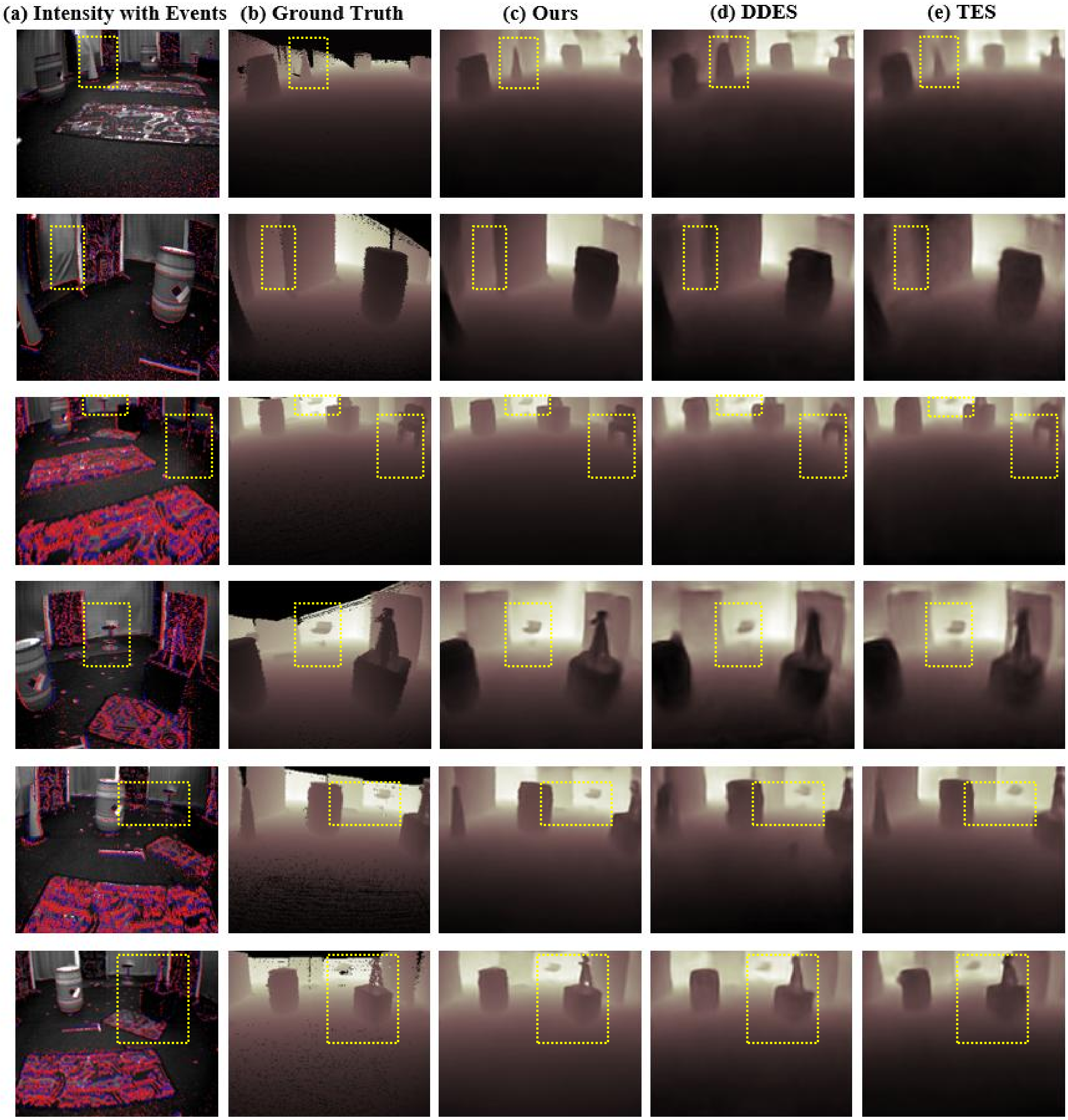

Figure 2. More qualitative comparison of dense disparity estimation for indoor flying scenes in the MVSEC dataset. (d) and (e) are the results of DDES [4], and TES [2] respectively. Detailed disparity information is framed by the yellow box for comparison.